Sentences style classification in russian language

Daniil Devyatkin, Andrew Tkachenko

December 2022

Abstract

This paper presents a solution of the classifying sentences and texts problem according to five functional styles of speech: scientific style, official business style, journalistic style, colloquial style, artistic style. Two methods were considered to solve this problem: the method of classical machine learning and the method based on deep learning. It was possible to achieve high classification quality indicators on a relatively small data set as a result of the thorough work. `https://github.com/d3vyatk4ru/style-sentences-clf`.

1 Introduction

When studying some topics on the subject “Russian language” in the primary grades of schools, it happens that there are not enough prepared materials for teachers, methodological guides are copied from year to year, and examples in students’ assignments do not shine with variety. An example of such a topic is "functional styles of speech", which involves determining the style of a presented text or sentence. This work is aimed at solving such a problem using machine and deep learning tools. The work involves several logical stages: collecting and marking the dataset, cleaning and preprocessing the dataset, choosing the best model that solves this problem based on a comparative analysis. Popular sources of information were used to create a labeled dataset: news sites, literary works, an archive of scientific publications and articles, etc. As a result of training models on the collected dataset, interesting results were obtained, which will be demonstrated in subsequent chapters.

1.1 Team

Andrew Tkachenko (tkachenko_aa10@mail.ru) collect and prepared dataset, prepared this document.
Daniil Devyatkin (danya.devyatkin@mail.ru) realization and train ML / DL models, prepared this document.

2 Related Work

In this section, you will describe in details the existing approaches to the problem you work on. For each approach, you need to provide a reference.

[?] is a sample reference to the previous art. [?] is a sample reference in Russian.

3 Model Description

Two models, which are based on different approaches, were implemented in this work: classical machine learning and deep learning method. The TF-IDF approach was used, when using classical machine learning for feature extraction. The TF-IDF algorithm calculates the importance of a word for a document relative to other documents. The main hypothesis is that if a term is used frequently in a certain text but rarely in others, then it is of greater significance for that text. The formalization of the method is as follows:

$$TF(w, D) = \frac{count(w)}{|D|},$$

$$IDF(w, d) = \log \frac{|D|}{\{d \in D \,|\, w \in d\}},$$

$$TFIDF(w, d, D) = TF \cdot IDF,$$

where d - document, w - words in document d, D - document collection.

Logistic regression was used as an algorithm for solving the multiclass classification problem.:

$$P(y = j \,|\, x) = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}},$$

$$Q(X, w) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} [y_i = j] \log p(y = j \,|\, x_i).$$

Main steps:

- text cleaning;
- token extraction using TF-IDF approach;
- using logistic regression to predict text style.

Another approach was based on deep learning methods. The word2vec approach was used to search for word embeddings. Training was not carried out

from scratch, the Gensim library was used with pre-trained word representation models. A recurrent neural network was used as a model, and to be more precise, the GRU model (Fig. 1):

$$z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z), \tag{1}$$
$$r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r), \tag{2}$$
$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h), \tag{3}$$

where x_t - input vector, h_t - hidden state vector, z_t - renovation valve vector, r_t - relief valve vector, W, U - parameter matrices, $\circ$ denotes the Hadamard product.

To predict the class label a fully connected layer of size 5 was used as the last layer. Cross entropy was used as the loss function.

Main steps:

- tokenization;
- use of pre-trained vector representations for Russian words;
- latent state calculation in the recurrent network;
- sentence type prediction.

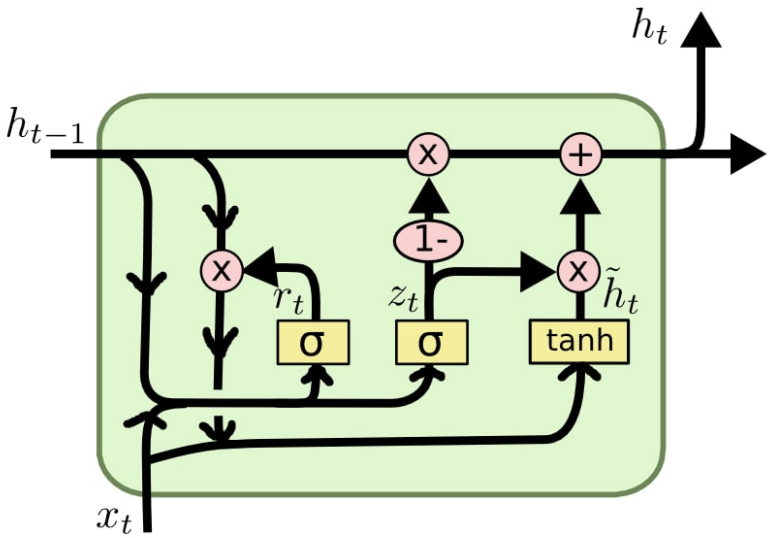

Figure 1: GRU model architecture.

4 Dataset

The entire dataset, which was necessary for work, was assembled independently and consists of 2500 elements. Examples of texts and sentences of different styles were taken from works of art of different times, news materials on different topics, scientific articles and publications, messages and comments from open resources and social networks. Analogues of such a dataset were not found on the network.

5 Experiments

This paragraph describes the metrics used to evaluate the quality of the models, the setup of the models and dataset used, and the baseline that was targeted.

5.1 Metrics

Let us introduce some metrics for evaluating the quality of models. The classification error matrix is the matrix:

$$M = \{m_{ij}\}_{i,j=0}^{N},\ m_{ij} = \sum_{k=0}^{N} \mathbb{I}\left[\hat{y}_k = j\right] \mathbb{I}\left[y_k = i\right], \tag{4}$$

where N — number of classes. On the main diagonal are correctly predicted objects. The error matrix for binary classification is presented in the table.
!!!!!!!!!!!!!!!TABLE!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!TABLE!!!!!!!!!!!!!!!
!!!!!!!!!!!!!!!TABLE!!!!!!!!!!!!!!!
Based on the error matrix, we introduce the F1 metric, which is calculated using auxiliary metrics. Let's describe them. To calculate the F1 metric, you need to know the accuracy metric values (*recall*) и полнота (*precision*):

$$recall = \frac{TP}{TP + FN}, \qquad precision = \frac{TP}{TP + FP}.$$

Value of *recall* metric shows, what proportion of class objects +1 of all objects of class +1 the model correctly predicted. In other words, *recall* characterizes classification errors — the larger the value of *recall*, the more correct predictions. The *precision* metric is interpreted as the proportion of objects predicted by the classifier as class +1, and at the same time, indeed, belonging to class +1. For the joint evaluation of *recall* and *precision* use the F metric. F metric — harmonic mean between *precision* and *recall*:

$$F = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}, \tag{5}$$

where β — a weighting factor that allows you to increase the contribution of *precision* or *recall*.For $\beta < 1$, preference is given to the *precision* metric, and

for $\beta > 1$, the *recall* metric is preferred. If $\beta = 1$ is substituted into (5), then both metrics contribute equally to the estimation of the model prediction quality, and the F1 metric is obtained:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

5.2 Experiment Setup

Several hyperparameters varied when training a model based on the classical approach: random seed in splitting the dataset into test and training samples, the logistic regression regularization parameter, the optimization problem solver, and the number of iterations. The best result was shown with the following parameters: solver - liblinear, logistic regression regularization parameter - 1000, number of iterations - 2000, random seed - 31.

There were more parameters to vary for the deep learning approach: number of epochs, hidden state vector in GRU, number of layers in RNN, dropout probability, optimizer, training step, batch size, etc.

5.3 Baselines

The "coin toss" case was considered as a baseline, : that is, the probability of getting the correct class label for each object is 0.2.

6 Results

It was noted earlier that two approaches were used in the work. When using the TF-IDF and logistic regression approach, we managed to achieve a score on the F1 metric equal to 0.89. This value is an average of 100 trainings. Also in Fig. 2 demonstrates the error matrix for this approach.

The matrix has a pronounced diagonal. This demonstrates a fairly accurate performance of the model, but there are also classification errors. Tab. 2 shows an example of correct and erroneous classification of texts.

data	ground truth	predict
Compulsory health insurance - type of compulsory social insurance, which is a system state-created legal, ...	0	1
However, some atoms lose their electrons or capture others when connected to each other	0	0

Table 1: Output samples.

Let's move on to the recurrent neural network approach. Here, the quality of the network in F1 measure is 0.954, which is not surprising, since the deep learning approach has a greater generalization ability. Also on Fig. 3. error matrix is presented.

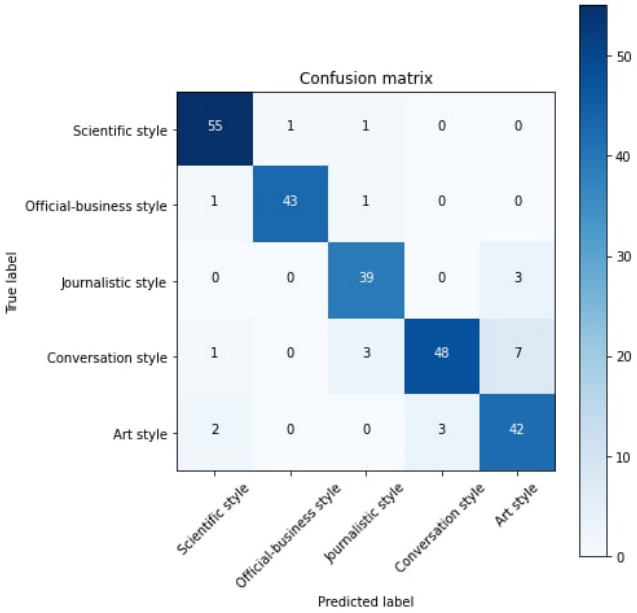

Figure 2: Confusion matrix for classical ML approach.

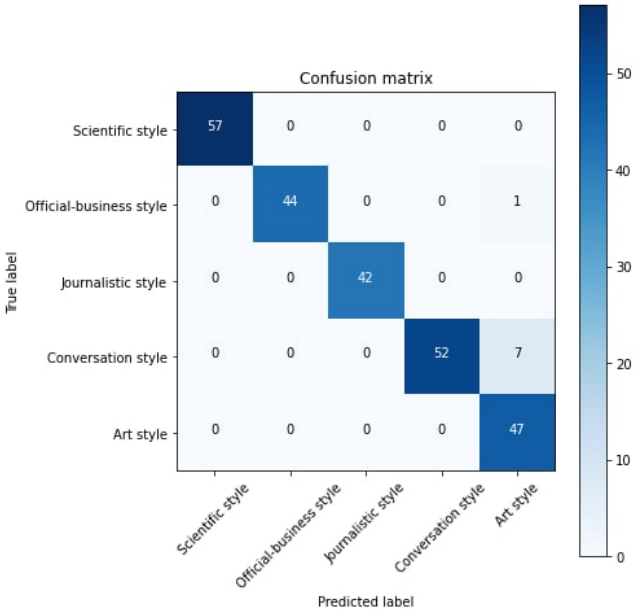

Figure 3: Confusion matrix for DL approach.

data	ground truth	predict
God! Negro is made by a man!	3	2
this year on the only Russian inland antarctic station east start construction of a new building	2	2

Table 2: Output samples.

7 Conclusion

The following was done to implement this project: a self-assembled dataset, consisting of 2500 elements, was marked up and preprocessed; a comparison of two different approaches based on classical machine learning methods and deep learning methods was presented. As a result of the work, the superiority of the deep learning method for solving this problem was revealed, in particular, due to its greater generalizing ability.